Discovering Potential Founders Based on Academic Background

Arman Arzani[1][a], Marcus Handte[1][b], Matteo Zella[2][c] and Pedro José Marrón[1][d]

[1] *University of Duisburg-Essen, Essen, Germany*

[2] *Niederrhein University of Applied Sciences, Krefeld, Germany*

Keywords: Knowledge Transfer, Founding Potential, Researcher Profiling, Innovation Identification.

Abstract: Technology transfer is central to the development of an iconic entrepreneurial university. Academic science has become increasingly entrepreneurial, not only through industry connections for research support or transfer of technology but also in its inner dynamic. To foster knowledge transfer, many universities undergo a scouting process by their innovation coaches. The goal is to find staff members and students, who have the knowledge, expertise and the potential to found startups by transforming their research results into a product. Since there is no systematic approach to measure the innovation potential of university members based on their academic activities, the scouting process is typically subjective and relies heavily on the experience of the innovation coaches. In this paper, we study the discovery of potential founders to support the scouting process using a data-driven approach. We create a novel data set by integrating the founder profiles with the academic activities from 8 universities across 5 countries. We explain the process of data integration as well as feature engineering. Finally by applying machine learning methods, we investigate the classification accurracy of founders based on their academic background. Our analysis shows that using a Random Forest (RF), it is possible to successfully differentiate founders and non-founders. Additionally, this accuracy of the classification task remains mostly stable when applying a RF trained on one university to another, suggesting the existence of a generic founder profile.

1 INTRODUCTION

Universities play an important role in adding social impact through teaching and education. Also, their interaction with industry is essential to innovation and to a knowledge-based economy. While universities dominate the principle of knowledge-based communities, industry represents the primary institution in industrial societies, therefore remaining a key factor as a locus of production. By comparison, one crucial advantage of universities over industry as knowledge-producing institutions, is the cluster of students, graduates and post-graduates. While industrial research and development (R & D) units of government and firm laboratories tend to solidify over time due to the lack of continuous flow of human capital, the universities profit greatly from it.

Today, many universities are extending their traditional role from education and research towards research transfer. Research transfer is the joint development and dissemination of knowledge as a product that has social contributions such as sharing, communication of experience, building contacts and innovation networks. According to surveys, 55% of spin-offs draw on tacit knowledge acquired at the university, whereas only 45% use codified research findings from the university (Karnani, 2013). Moreover, research transfer also capitalizes on the knowledge and human base, by conveying research results to a broader audience, which is particularly useful for people who want to start their own companies.

The universities that practice research transfer want to support potential founders, start-ups and innovation. Providing effective support requires the universities to identify the potential founders within their organization. In many universities, this scouting process is done by innovation coaches. As part of the process, the innovation coaches manually monitor the research activities at their university and conduct interviews. Since there is no systematic approach to measure the innovation potential of university members based on their academic activities, the scouting process is typically subjective and relies heavily on the experience of the innovation coaches.

In this paper, we study the problem of discovering potential founders to support the scouting pro-

[a] https://orcid.org/0009-0000-1304-9012

[b] https://orcid.org/0000-0003-4054-1306

[c] https://orcid.org/0000-0003-1830-9754

[d] https://orcid.org/0000-0001-7233-2547

Arzani, A., Handte, M., Zella, M. and Marrón, P.
Discovering Potential Founders Based on Academic Background.
DOI: 10.5220/0012156200003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 3: KMIS*, pages 117-125
ISBN: 978-989-758-671-2; ISSN: 2184-3228

cess using a data-driven approach. We create a novel data set by integrating the founder profiles of Crunchbase with the academic activities using Dimensions. Thereby, we generate large data set with more than 11.000 founders and non-founders from 8 universities across 5 countries that encompasses more than 4 million publications, 3 million patents and 80.000 grants including the relations among them.

While most of the related work focuses on either success rate prediction of startups, i.e., venture capital prediction or researcher profiling inside the university, we apply machine learning methods to support the discovery of potential founders.

The main contributions of our paper are:

- Creating a novel dataset by joining Crunchbase and Dimensions data sources, that contains founders as well as non-founders with their academic metadata
- Extracting features based on the academic meta-data and performing classification on the founders/non-founders dataset
- Providing quantitative evidence for the existence of a generic founder profile across multiple universities and countries

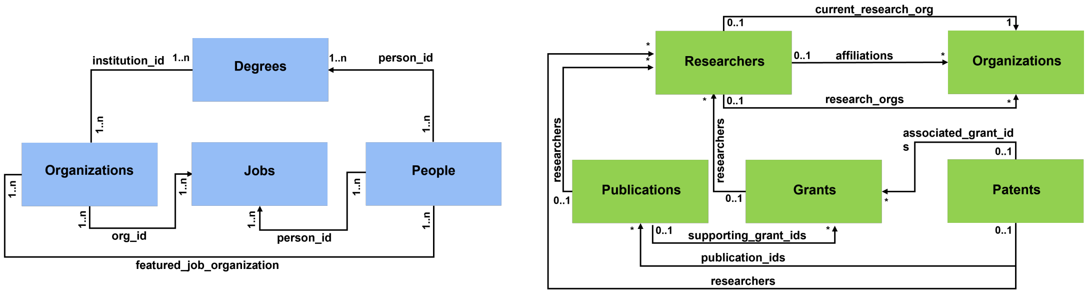
2 BACKGROUND

Academic entrepreneurship has played an important role in fostering regional economic development and defines the process of commercialization of science at the universities, as well as other patterns of technology transfer that focus on licensing, patenting and start-up activity. Moreover, universities around the world increasingly encourage the involvement of their academics in the transfer of knowledge to the marketplace through spin-off activities which enhances economic growth and regional competitiveness (Siegel and Wright, 2015, González-Pernía et al., 2013).

In order to support reproducibility in knowledge transfer, many researchers have studied the process of creating university spin-offs (Pirnay et al., 2003, Van Burg et al., 2008). Authors of (Pirnay et al., 2003) suggested a topology for university spin-offs which was lacking depth at the time by underlying studies. In their work they presented a two-dimensional system that stimulates academic spin-offs, including the status of the person and the area of knowledge. Additionally, to further understand underlying processes in academic spin-offs and improve the performance of the incumbent university, the authors of (Van Burg et al., 2008) described a case study of Eindhoven University of Technology. They introduced and argued on how two design principles, namely research-based (tacit knowledge of conversion of key agents in university) and practice-based principle (practices and experiences of the agents) stimulates academic spin-offs. Both of the previous works are limited to sketching the theoretical foundation of academics and apply empirical studies.

Other than defining the general structural backbone of academic spin-offs, some related work also focused on specific factors of successful academic entrepreneurship (Müller, 2010, Backes-Gellner and Werner, 2007). Authors of (Müller, 2010), focus on analyzing time lag of startup creation. By providing empirical results through a survey, they argued that spin-offs are done years after the academics had left the academic institutions, since the founders need time to gather practical experience by working in industry. This was contradicted by (Backes-Gellner and Werner, 2007) that investigated the educational signals of academics. Their evidence showed that the academics who finished their university degree faster than others and have patents, have a higher chance of starting their venture. Furthermore, some works suggested that the innovation capabilities and systems are different across countries and regions (Wright, 2007, Rothaermel et al., 2007). While this might be true in general, the results of our quantitative evaluation suggest that the founder profile remains mostly stable for the (international) set of universities covered by our data.

Most machine learning methods regarding entrepreneurship and startup success revolve around entrepreneurial finances and the prediction of the success rate for the startup phase (Ferrati et al., 2021,Sharchilev et al., 2018,Żbikowski and Antosiuk, 2021). Only few related works deals with the identification of entrepreneurs in various context (Montebruno et al., 2020, Chung, 2023) using machine learning. Based on data from historical British entrepreneurs, the authors of (Montebruno et al., 2020) proposed a model that was able to classify employment status and could identify entrepreneurs from workers. This work used self-reported data in the Victorian censuses over 1851-1911 and investigated whether a trained classifier on later censuses can identify entrepreneurs in the early censuses using features such as age, district, marital status, number of servants, etc. Following this approach, the authors of (Chung, 2023) developed a classification pipeline based on the data of the Global Entrepreneurship Monitor. The study included survey data on adult population and their entrepreneur characteristics, as well as their social attitudes towards entrepreneurship. The authors also considered primary individ-



(a) Data model of Crunchbase.

(b) Data model of Dimensions.

Figure 1: Crunchbase and Dimensions data sources.

ual characteristics such as age, gender, education as well as environmental attributes to train classifiers that could identify potential entrepreneurs. In another study (Sabahi and Parast, 2020), the authors developed classifiers to predict an individual's project performance based on entrepreneurial features, such as founding attitude, social self-efficacy, appearance self-efficacy, and cooperativeness, and entrepreneurial orientation such as proactiveness.

Whereas previous studies depend on surveys and empirical results from individual institutions, our work takes academic activities into consideration, since our primary focus relies on improving knowledge transfer in the context of innovation scouting at universities. In addition, our approach is automatable since the data on academic activities of staff members such as bibliometrics and scientometrics are often already collected by universities for other purposes. To our knowledge, none of the related work in knowledge transfer have tried to tap into the academic information of researchers to identify existing entrepreneurial activities and have considered the use of academic features in investigating the founder's profile. In this paper, we propose a novel and automatable approach to support innovation coaches, that uses machine learning classifiers to identify potential entrepreneurs inside universities. To apply machine learning, we introduce a set of features derived from the academic activities of the staff members, including their publications, patents, and grants as well as their impact. This allows us to take advantage of bibliometric and scientometric data of the researchers, which is often already available or can be gathered from online data sources.

3 DATA SET

To study the discovery of potential founders in universities based on their academic activities, we use Crunchbase (crunchbase.com, 2007) and Dimensions (Hook et al., 2018) as our data sources. In the following, we first provide details on both data sources. We then explain, how we integrate their data into a single data set consisting of founders and non-founders. Thereafter, we introduce a set of features, which we extract for further analysis and outline the reasoning for choosing them. Finally, we describe the concrete details of the data generation process resulting in the data that is used for further analysis.

3.1 Data Sources

3.1.1 Crunchbase

To gather information about founders, we received access to the Crunchbase database[1]. Crunchbase (crunchbase.com, 2007) is a data as a service platform with business information about private and public companies, founders, or people in leadership positions, investors and founding rounds. Crunchbase was originally a database to track the startups featured in the TechCrunch website[2]. At the time of writing, it encompasses more than 2 million organizations. The database consists of multiple tables that can be joined by unique identifiers. A simplified entity-relationship diagram (ERD) is shown in Figure 1a.

The organizations table includes information about companies as well as investors and universities which are differentiated by type. The table contains fields such as name, address, number of employees and the status of the organization (active, closed, acquired). The people table describes individuals who are founders, investors, or employees of one or more organization(s). This table includes the person's name, gender, address, social media account links, organization, and job position within the organization. The job position belongs to the jobs table as well. The information about an individual's educational background is held in the degrees table. Each

[1] https://data.crunchbase.com/

[2] https://techcrunch.com/

record contains details about the subject of the acquired degree, date of matriculation and graduation, as well as the institution awarding the degree. The institutions are addressable through the organizations table.

Since we are interested in analyzing the startup spin-offs by academics, we perform an exploratory data analysis (EDA), by joining the degrees, organizations, and people tables. Using the organization type, we filter the organizations to only select the universities. To only focus on founders, we utilize the job type from the jobs table to exclude employees as well as other operational job titles. At this stage, by joining people, degrees and organizations and adjusting the query on a specific university, we can extract the information of the founders who have studied or are currently studying at a specific university.

3.1.2 Dimensions

To gather information about academic activities, we received access to Dimensions. Dimensions (Hook et al., 2018) is a modern data infrastructure for discovery and research. It provides access to over 2.9 million grants, 121 million publications, citations and 140 million patents among others.

The publications of the Dimensions database consist of journal articles, pre-prints, books/book chapters with full text search available through their API access for more than 160 publishers (PubMed, arxiv, Crossref, etc.). Furthermore, the publications are highly contextualized with linked related grants, publication references, citing publications and related patents as shown in Figure 1b. This is a significant difference of Dimensions in comparison to its counterparts such as Scopus or Web of Science that makes it possible to analyze the relationship of patents, publications, or grants that reference each other at some point in time. Furthermore, the grants hold project funding in the private, federal as well as national sector which are either crawled directly from the funders websites or extracted through their APIs. The patent data covers over 100 jurisdictions, with the information containing inventions, bibliometrics and the original institutions that funded the patents.

To access the data, Dimensions provides a rich set of APIs for full-text queries that we primarily used to perform keyword searches for all the instances of a term in a document or group of documents. Using the full text API of Dimensions, specific sections of a document, such as abstract, full text, authors can be targeted and due to the links between different data types, it is possible to create complex filters.

3.1.3 Data Integration

With data integration we pursue two goals: First, we need to connect the data on founders available in Crunchbase with the academic activities of each individual founder in Dimensions. Second, since we want to apply machine learning for founder discovery, we also require corresponding non-founders data in a similar quantity as the founder data.

To fulfill the first goal, we start with the extracted founder records from Crunchbase that include the founder's name, degree and university name. Using the name of a founder, we could try to identify relevant documents by running full-text queries against the Dimensions API and collecting the results with matching author names. However, the results of such queries would likely contain many false positives, since person names are often not unique. In practice, this problem is further amplified by the large scale of Dimensions and by very common family names such as Xu or Zhu.

To mitigate this problem, we take the founder's university into account. To do this, we start by mapping relevant organizations of Crunchbase to organizations in Dimensions. To identify organizations, Dimensions relies on the unique organization IDs of the Global Research Identifier Database[3] (GRID). GRID is a free online database that provides information about research organizations and addresses the problem of messy and inconsistent data on research institutions, ensuring that each entity is unique. GRID stores the type of institution, geo-coordinates, official website, Wikipedia page and name variations of institutions for each ID and offers an online search tool to lookup IDs by name. Since the data volume is low, we perform the mapping manually using the online tool.

Given the GRID ID for organizations in Crunchbase, we can refine the full-text queries issued to Dimensions to only include matches of researchers that exhibit the desired ID. However, since researchers may change their jobs over time, the same person may belong to multiple organizations over time. Taking this into account, we formulate the queries such that they select matching person names whose organizations include the target ID.

While this greatly reduces the number of false positives, it does not prevent them. An example for this would be two persons with the same name that worked at some point in their careers at the same university. To eliminate such cases, we gather the full dataset for each university and then perform a statistical outlier detection on the data to remove the resulting anomalies. Since the problem can only gener-

[3] https://www.grid.ac/

Table 1: Extracted and generated features for model training.

Category	Name	Feature	Description	Type	Dimensions	Crunchbase
Person	Name	full_name	Person's name	String	✓	✓
	Founded	is_founder	Whether the person is founder	Boolean	×	✓
Publication	Publications	pub_count	Number of publications	Integer	✓	×
	Mean citation	pub_citation_mean	Average citations	Float	✓	×
	I10-index	i10_index	I10 index of citations	Integer	✓	×
	H-index	h_index	H index of citations	Integer	✓	×
	G-index	g_index	G index of citations	Integer	✓	×
	Industrial research	industry_collab_research	Number of papers with industry	Integer	✓	×
	Innovation impact	research_innovation_impact	Number of linked patents	Integer	✓	×
	Affiliations	org_affiliation	Number of institutional affiliations	Integer	✓	×
Patent	Patents	pat_count	Number of patents	Integer	✓	×
	Mean citation	pat_citation_mean	Average citations	Float	✓	×
Grant	Grants	grant_count	Number of Grants	Integer	✓	×
	Research impact	research_impact	Research output of grants	Integer	✓	×
	Grant innovation impact	grant_innovation_impact	Number of linked patents	Integer	✓	×

ate too much data, we can focus the outlier removal on cases with an unlikely high number of data items. While this process reduces the data quantity by reducing the set of founders, it ensures that the quality of the remaining data stays high. Given that the outlier removal generally affects less than 4% of the data, we think that this process is a reasonable trade-off.

To address our second goal of generating a set of non-founders for each organization, we can build upon the mapping of Crunchbase organizations and GRID IDs. As a first step, we use the GRID ID of a university to generate a list of researchers belonging to the organization in Dimensions. Thereafter, we randomly pick researchers that are not contained in the Crunchbase database for the organization, and we start issuing the same queries against the Dimensions API as for the founders. The resulting datasets include university-specific founders and non-founders records along with their publications, patents, as well as grants information and their linked metadata.

3.2 Feature Extraction

The data integration of Crunchbase and Dimensions generates sets of founders and non-founders from different universities together with their academic activities, i.e., publications, patents and grants. We hypothesize that the latent founding potential lies beneath the academic profile of the founders and can be transformed into a predictive model. Since we already know whether the persons have founded a company or not, we can apply supervised machine learning algorithms to train a classifier. For the classification to be effective, we also need to identify a set of features that we can extract from the data and that might be useful to differentiate founders and non-founders.

The data available for each person can be classified into four categories, namely person, publication, grant, and patent. Table 1 shows a detailed listing of the features extracted for each category including a feature name, description, data type and origin.

In the person category, we use Crunchbase to extract founder information that contains the *name* of the founder. For non-founders, we only store the person's name which we get from Dimensions and only use the *founded* feature that differentiates non-founders and founders.

In the publication category, we extract the number of *publications* as a basic metric for researcher productivity. However, since knowledge transfer is a cumulative process that happens through research and practical work, we also want to capture the importance, significance, and broad impact of a scientist's cumulative research contributions. To do this, we extract the citation information for each publication and derive several scientometric features, including the average number of citations, the I10-index, the H-index and the G-index. Although these features share the common goal of quantifying the research productivity and impact, each of these features emphasizes a different aspect. For example, the I10-Index, which is used by Google Scholar, only counts publications that received a minimum of 10 citations. In contrast to this, the H-index (Hirsch, 2005) stresses the importance of citations, since a researcher has the index h, if their n paper has at least h citations each and the other n papers have no more than h citations. The G-index is an even more developed version of H-index and aims to improve H-index by giving more weight to highly-cited papers. Given a set of publications ranked in decreasing order of the number of citations that they received, the G-index is the (unique) largest number such that the top g articles (together) received at least g squared citations (Bihari and Pandia, 2015).

In addition to productivity and overall impact in the research community, we also want to capture the relevance of publications to the industry. To do this, we extract the number of publications that are linked to patents as a way to estimate the *innovation impact* and count the number of papers published in collaboration with industrial partners as a way to identify *industrial research*. Finally, as our last feature in the publications category, we count the number of *affili-*

Table 2: Numbers and statistics of the extracted data.

			Founder(Company)				Non-founder			
Institute	**Country**	**Abbrev.**	**Total**	**Publication**	**Patent**	**Grant**	**Total**	**Publication**	**Patent**	**Grant**
Stanford University	U.S.	SU	1785	400719	299706	8571	1785	1265472	940805	16784
University of California, Berkeley	U.S.	UCB	1193	421114	261916	7690	1193	619310	463897	13039
Harvard University	U.S.	HU	1065	356078	223908	6776	1065	573964	452765	8994
University of Oxford	England	OU	597	179273	32670	4332	597	257776	219040	4587
Tel Aviv University	Israel	TAU	626	27771	25841	485	626	162934	6341	3069
University of Toronto	Canada	UofT	421	95995	81516	3009	421	195503	176289	4205
Technical University of Munich	Germany	TUM	244	29020	21313	759	244	89064	80692	1304
University of Duisburg-Essen	Germany	UDE	20	875	557	13	20	5791	646	105

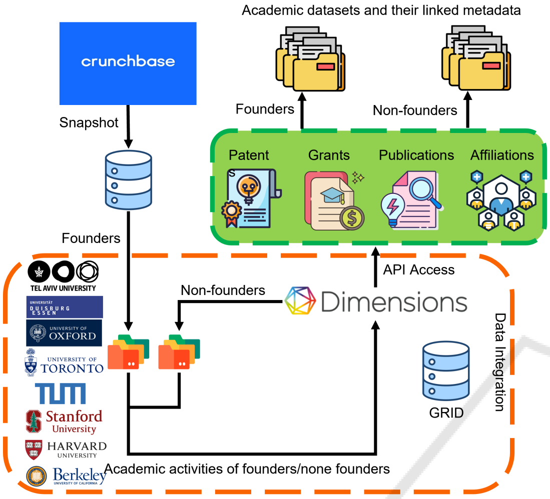


Figure 2: Data generation pipeline.

ations of each person as a simple metric to track the person's career path.

With the patent category, we try to capture the practical applicability of academic activities as basis for innovations. There, we simply extract the number of *patents* and in addition, we calculate the *mean citation* of the patent documents to estimate the potential impact of a researcher's patent portfolio.

Finally, the grants category can be seen as a tool for fostering knowledge acquisition and transfer. For this category, we first extract the number of research *grants*. For each grant, we then determine the *research impact* by extracting the number of scientific publications linked to the grant and then we determine the *grant innovation impact* by extracting the number of linked patents. This completes the triangle of publications, patents, and grants and embeds their influence on each other.

3.3 Data Generation

To generate the data set, we implement the data integration and feature extraction logic described previously as a processing pipeline using Python. The general flow through the pipeline is depicted in Figure 2. When started, the pipeline first downloads a snapshot of Crunchbase in CSV format and generate an SQLite database. Thereafter, it extracts the sets of founders for a set of target organizations and then it generates a set of non-founders using Dimensions. Using the resulting list of persons, the pipeline issues queries against the Dimensions API to retrieve the data required to compute the features. Due to the high number of queries, we locally store their results so that re-executing the pipeline will not cause duplicate API calls. Once the data is available, the pipeline computes the features for each person and performs the outlier detection and removal described previously. The result are two datasets for each organization that contains the features for founders and non-founders.

To generate the data used for the analysis, we use the daily snapshot of the Crunchbase database from October 18, 2022. We pick eight universities across the US, England, Israel, Canada, and Germany. Together the universities cover the whole size spectrum with respect to the total number of founders with Stanford University being among the universities with the largest number of founders and the University of Duisburg-Essen being among the lowest ones. The goal here is not to necessarily pick top universities. Instead, based on the exploratory analysis, we tried to pick universities with a varying number of founders to get more meaningful results when investigating the differences between founders and non-founders. The data set also includes some universities that exhibit a lower ranking such as Duisburg-Essen and Toronto. In total, this selection results in 5951 founds which we augment with an equal-sized set of non-founders using Dimensions. Using the Dimensions API, we download the academic activities for the resulting 11902 researchers. In total, this results in 4,680,659 publications, 3,287,902 patents, 83,722 grants including their linked metadata as depicted in Table 2.

4 DATA ANALYSIS

Given the generated data set described in the previous section, we utilize data analysis to answer the two questions: First, is it possible to accurately classify founders and non-founders using features derived from their academic activities? Second, are there sig-

Table 3: F1 Scores for decision tree (DT) and random forest (RF): Training with one institute and validation using others.

Training \ Validation	SU		UCB		HU		OU		TAU		UofT		TUM		UDE	
	DT	RF	DT	RF	DT	RF	DT	RF	DT	RF	DT	RF	DT	RF	DT	RF
SU			0.73	0.78	0.75	0.76	0.74	0.77	0.83	0.83	0.73	0.76	0.71	0.74	0.74	0.77
UCB	0.76	0.78			0.72	0.73	0.72	0.77	**0.86**	0.84	0.73	0.75	0.71	0.72	0.74	0.77
HU	0.78	0.79	0.75	0.78			0.74	0.78	0.82	0.81	0.75	0.75	0.71	0.76	0.82	0.85
OU	0.77	0.79	0.76	0.77	0.74	0.75			0.80	0.82	0.72	0.75	0.75	0.76	0.85	**0.87**
TAU	0.73	0.71	0.69	0.69	0.68	0.66	0.70	0.66			0.71	0.69	0.76	0.76	0.85	0.77
UofT	0.75	0.77	0.74	0.75	0.71	0.73	0.71	0.74	0.84	0.80			0.71	0.71	0.77	0.79
TUM	0.74	0.74	0.70	0.72	0.71	0.72	0.71	0.71	0.84	0.85	0.68	0.69			0.79	0.85
UDE	0.61	0.74	0.58	0.71	0.57	0.71	**0.56**	0.70	0.62	0.78	0.58	0.69	0.58	0.75		

nificant differences between the founders at different universities? To this end, we address this question using machine learning methods which can recognize patterns based on the input data. To implement the analysis, we rely on Python and use the Pandas and Scikit-learn (Pedregosa et al., 2011) libraries.

With classification, we first determine whether it is possible to accurately classify founders and non-founders using features derived from their academic activities. To do this, we use the features contained in our data set (c.f. Table 2) as an input to generate a classifier using a machine learning algorithm.

To decide on the type of machine learning algorithm, we considered generative as well as discriminative algorithms. Generative algorithms such as Naive Bayes work under the assumption that no feature correlation exists. This assumption, however, does not hold for our data set. For example, the higher the number of publications, the higher are the number of citations in most cases. For this reason, we choose discriminative algorithms.

Since the explainability of the results and model's simplicity play a significant role in the context of this paper, we choose Decision Tree (Kotsiantis, 2013) (DT) and Random Forest (Breiman, 2001) (RF) classifiers. Benchmark studies have demonstrated that RF and DT classification algorithms are among the best classifiers for many real-world datasets (Fernández-Delgado et al., 2014, Olson et al., 2017) and both are conveniently interpretable. While DTs work with a single tree, RFs avoid and prevent overfitting by binding multiple trees. This often results in a higher accuracy and more generalizable models.

For several machine learning algorithms such as regression tasks and neural networks, it is necessary to bring the range of all numerical variables to a common scale. This ensures that each feature will receive an equal importance during the time of training. However, for DT and RF this type of scaling is not necessary since they do not compute distances between features but rather identify thresholds in individual features. As a split scoring function, we use entropy information gain for both DT and RF classifiers, therefore normalization is not required (Li and Zhou, 2016). In addition, by preserving the scale of the numerical values we can analyze the partitions of the raw features to have a better image of the quantity of involved features in founding potential. For example, it would be easy to answer whether a specific number of publications or patents are needed to have a higher founding potential.

Before training, we apply hyperparameter tuning, to reach a higher accuracy while avoiding model overfitting. To do this, we run a grid search on a subset of predefined parameters. For decision trees we estimate maximum number of levels, minimum number of samples for node splitting and minimum number of samples for a leaf node. The grid search for random forests parameters also encompasses these parameters but in multiple trees. In addition, we determine the estimated number of trees and whether bootstrapping is needed. Bootstrapping is the process of randomly sampling subsets of a dataset over a given number of iterations and a given number of variables. These results are then averaged together to obtain the final result. The best hyperparameters are selected from the grid search and passed to the corresponding classifier. Finally, to ensure that the accuracy scores are robust, we perform a 10-fold cross validation (CV).

As a first step, we take 50 percent of the data of each university for the classifier training and we use the remaining 50 percent of the data to compute an accuracy score. While splitting the dataset, we apply stratified sampling to ensure that the number of founders and non-founders remains balanced in both, the training and the validation set. After hyperparameter tuning and with a 10-fold CV, this process yields an F1 score of 76% for the DT classifier. For the RF, the F1 classification accuracy is slightly higher with 79%. Given these scores, we can conclude that features extracted from the academic activities can indeed be used to properly identify a significant number of founders. However, we also note that the classification accuracy is not perfect. Given that innovative research results are probably not the only relevant decision criteria when founding a company, the imperfect outcome does not seem to be overly surprising. To clarify this consider that there are several other factors such as family wealth, risk disposition or country's economic situation that also play an important role during a career decision. Since these factors are not captured by the academic background, we would assume that our dataset cannot explain all differences between founders and non-founders. To further test this assumption, we have experimented with a number of more advanced classification techniques such as XGBoost and feed-forward neural networks. The accuracy of these approaches is generally similar or worse.

Given the high accuracy scores for the identification of founders at individual universities, we continue the analysis by determining the sensitivity of the classification model with respect to the university. To do so, we take the full dataset of each university as an input to train a classifier. Thereafter, we apply this classifier to the full dataset of all other universities. Given that we are training a DT as well as a RF, this results in 112 (2*8*(8-1)) accuracy scores of 16 classifiers. Table 3 shows the results and highlights the highest and lowest scores for the DT and RF classifier.

On average, this experiment results in an accuracy score of 73% for the DT and 76% for the RF. Overall, the worst classification performance is yielded by the classifier trained on the data of the UDE which also exhibits the lowest performance (56% for DT and 58% for RF). When looking at the performance of the other classifiers on the data of the UDE (last two columns), however, it becomes apparent that the low performance is rather an artifact of the low number of input values (20 founders and 20 non-founders) than a systematic difference between the UDE and other universities. To justify this, consider that the classifiers of other universities are able to classify the data of the UDE with an accuracy of at least 74%. Thus, it is safe to assume that the UDE data is simply insufficient to determine the proper thresholds during decision tree learning.

The remaining classifiers exhibit an accuracy between 66% and 87% with most scores lying around 75%. Given that the universities cover 5 different countries, we found this result to be surprising as it points towards the existence of a set of generic features that differentiate a significant number of founders from non-founders. Yet, despite the comparatively large number of founders and non-founders in our data set, due to the limited number of universities, we also think that further research is needed to harden or falsify this observation.

5 CONCLUSIONS

Academic science has become increasingly entrepreneurial, and many universities have started support programs to foster this type of technology transfer. An important goal of these programs is to find staff members and students that exhibit the knowledge and potential to transform their results into a product base. However, without a systematic approach to the discovery of potential founders, the scouting process can be challenging, and its success depends on the experience of the persons that execute it.

In this paper, we studied the discovery of potential founders using a data-driven approach. To do this, we created a data set that combines founder information with the corresponding academic activities and we applied machine learning methods to systematically study the data. Our analysis showed that it is possible to differentiate founders and non-founders with an average accuracy of 79%. This accuracy remains mostly stable when applying classifiers trained on one university to another, suggesting the existence of a generic founder profile.

At the current time, we are investigating the significance of the extracted features on the prediction of founded startups and study the impact of different research disciplines on the founder profile. Since our data sources also contain keywords for companies and research areas for academic activities, it would be interesting to determine whether (and how) the main discipline of a founder influences the founding potential or the resulting startup orientation. Thereafter, we are planning on building a graphical user interface around our data processing pipeline and analysis code. The goal is to make the system available to the innovation coaches of the science support center of our university. This will enable them to use the system as a tool to support their scouting process.

ACKNOWLEDGEMENTS

This work has been funded by GUIDE REGIO which aims to improve the ability of the science support center of the University of Duisburg-Essen in the identification, qualification, and incubation of innovation potentials. We thank Crunchbase and Dimensions for providing us with free access to their databases and APIs. This research would not have been possible without their generous support.

REFERENCES

Backes-Gellner, U. and Werner, A. (2007). Entrepreneurial signaling via education: A success factor in innovative start-ups. *Small Business Economics*, 29(1):173–190.

Bihari, A. and Pandia, M. K. (2015). Key author analysis in research professionals' relationship network using citation indices and centrality. *Procedia Computer Science*, 57:606–613.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chung, D. (2023). Machine learning for predictive model in entrepreneurship research: predicting entrepreneurial action. *Small Enterprise Research*, pages 1–18.

crunchbase.com (2007). Crunchbase: Discover innovative companies and the people behind them.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181.

Ferrati, F., Muffatto, M., et al. (2021). Entrepreneurial finance: emerging approaches using machine learning and big data. *Foundations and Trends® in Entrepreneurship*, 17(3):232–329.

González-Pernía, J. L., Kuechle, G., and Peña-Legazkue, I. (2013). An assessment of the determinants of university technology transfer. *Economic Development Quarterly*, 27(1):6–17.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572.

Hook, D. W., Porter, S. J., and Herzog, C. (2018). Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3:23.

Karnani, F. (2013). The university's unknown knowledge: Tacit knowledge, technology transfer and university spin-offs findings from an empirical study based on the theory of knowledge. *The Journal of Technology Transfer*, 38(3):235–250.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4):261–283.

Li, T. and Zhou, M. (2016). Ecg classification using wavelet packet entropy and random forests. *Entropy*, 18(8):285.

Montebruno, P., Bennett, R. J., Smith, H., and Van Lieshout, C. (2020). Machine learning classification of entrepreneurs in british historical census data. *Information Processing & Management*, 57(3):102210.

Müller, K. (2010). Academic spin-off's transfer speed—analyzing the time from leaving university to venture. *Research Policy*, 39(2):189–199.

Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., and Moore, J. H. (2017). Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, 10(1):1–13.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pirnay, F., Surlemont, B., Nlemvo, F., et al. (2003). Toward a typology of university spin-offs. *Small business economics*, 21(4):355–369.

Rothaermel, F. T., Agung, S. D., and Jiang, L. (2007). University entrepreneurship: a taxonomy of the literature. *Industrial and corporate change*, 16(4):691–791.

Sabahi, S. and Parast, M. M. (2020). The impact of entrepreneurship orientation on project performance: A machine learning approach. *International Journal of Production Economics*, 226:107621.

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P., and de Rijke, M. (2018). Web-based startup success prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 2283–2291.

Siegel, D. S. and Wright, M. (2015). Academic entrepreneurship: time for a rethink? *British journal of management*, 26(4):582–595.

Van Burg, E., Romme, A. G. L., Gilsing, V. A., and Reymen, I. M. (2008). Creating university spin-offs: a science-based design perspective. *Journal of Product Innovation Management*, 25(2):114–128.

Wright, M. (2007). *Academic entrepreneurship in Europe*. Edward Elgar Publishing.

Żbikowski, K. and Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using crunchbase data. *Information Processing & Management*, 58(4):102555.